Recovering the French Party Space from Twitter Data

Appendix

Contents

A Representativeness of the followers sample

A.1 Gender biais

By matching our full sample of Twitter followers with a dataset of first names (Campbell and New, 2000), we established that it over-represents males by a large margin, with approximately 2.2 males for one female among the 257,000 users (approximately a third of all followers) for which we could identify a unequivocal first name.

Furthermore, the average males-to-females ratio is even larger among the followers of each individual politician (at around 3 males for one female), and is larger for male politicians (at around 3.1) than it is for female politicians (at around 2.9), which suggests that the follower base of each politician might be even more dominated by males than the overall sample is.[1]

This last discrepancy, however, is corrected by our subsampling of Twitter followers, which brings down the average gender ratio for male and female politicians to 2.2 and 2.1 respectively, in line with the ratio observed at the level of the full sample. Therefore, while our sample does not adjust for the over-representation of males, it does not magnify it either.

A.2 Geographical bias

We located the followers by matching the 'location' field of their user profiles to extensive geographic information on all French administrative units and 470 of the largest cities. This approach successfully located over 60% of all users in France, but also certainly amplified our bias towards urban areas.

[1] The t-statistic for the difference in males-to-females ratios between male and female politicians was close to 3.7 ($p < 0.01$).

With regards to the geographical distribution of the sampled followers, Figure 1 shows the proportion of the sample located in each French metropolitan *département*, next to the proportion of the overall French population in those same units as of 2011. Unsurprisingly, those ratios are positively correlated (Pearson's correlation coefficient > 0.47 with $N = 96$ *départements*), which again suggests that our sample over-represents densely populated (urban) areas.

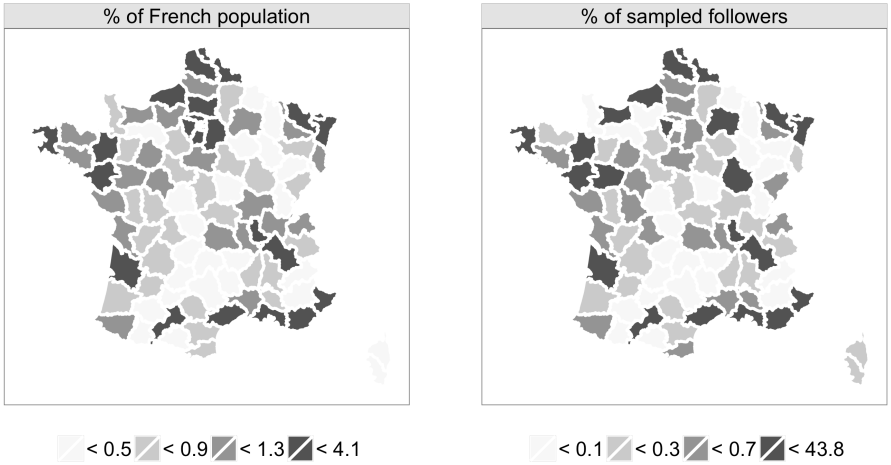

Figure 1: Geographical distributions of French population and followers sample, colored by percentage quartile. Census data by Insee (2012), map files by IGN (2015).

The *départements* in which we located a proportion of Twitter followers that is larger than the proportion of the population correspond to those with the largest cities, such as *Haute Garonne* (Toulouse), *Gironde* (Bordeaux), *Rhône* (Lyon), and of course, Paris, which contributes over ten times more Twitter followers to our sample (44%) than its population contributes to the French population (4%). As a consequence, our sample is only partially representative at the geographical level.[2]

We also looked at the correlation between the number of followers in each French *département* and the estimated population of five age groups in those same units

[2] Without Île-de-France, Pearson's correlation coefficient of both proportions increases to 0.77.

(Irdes, 2014). Interestingly, as shown in Table **??**, the geographical distribution of our Twitter followers best correlates with the distribution of the 20–39 age group, which might be due either to the general demographics of French Twitter users, or to the specific demographics of French Twitter users who follow politicians.

Age group	Correlation
0–19	0.36
20–39	0.55
40–59	0.44
60–74	0.44
75+	0.43

Table 1: Pearson's correlation coefficients between age-group population estimates and Twitter users at the level of $N = 96$ *départements*.

It should be noted that the unrepresentativeness of our sample is not a concern for our purposes, as we treat these followers as some form of an 'expert' survey of politicians' and their parties' positions (Barberá, 2015, p. 81). This point is well illustrated by the fact that the sample includes many privileged observers of politics, such as journalists from all kinds of French media.

B Estimation of partisan homophily

Figure 1 of the paper shows the unweighted, undirected graph of the adjacency matrix $M(i, j)$ of *i* politicians and their *j* followers on Twitter, collapsed to a one-mode network containing strictly politicians. The graph represents only a subset of that matrix, as a tie is set to exist between two politicians only when either one of them shares over half of his or her followers with the other one.[3] In the paper, we observe that the graph illustrates partisan homophily among Twitter followers, because it makes visually clear that politicians who are most likely to share a large proportion of their followers are those who also share a same party affiliation.

One way to estimate more precisely the amount of partisan homophily in this last graph consists in estimating the likelihood of a tie to exist between two politicians of the same party, controlling for the total number of ties in the graph and for other structural characteristics such as the propensity of ties to be reciprocal or transitive (that is, made through "friends-of-a-friend"). The network science literature offers an elaborate modelling strategy to estimate these parameters, known as exponential random graph models. In a nutshell, these models rely on simulations of the observed network to estimate the likelihood of tie formation, using an equation that can control for the kind of endogenous effects that emerge from network structures (Cranmer and Desmarais, 2011; Snijders, 2011).

As a means to illustrate the kind of findings that these models can provide about the data under study, Table 2 reports the results of an exponential random graph model of the network shown in Figure 1 of the paper. The equation of that model, which was written with the `ergm` R package (Hunter et al., 2008) and is available from the replication material, was set to estimate the propensity of politicians to

[3] From the viewpoint of network theory, collapsing a two-mode network to its one-mode representation is not necessarily recommendable, as it removes a great deal of structural information. Similarly, the threshold at which we chose to establish a tie is, in itself, arbitrary. Both steps, however, are heuristically useful at the exploratory stage, as they facilitate the visualisation of patterns of interest such as the one discussed here.

form ties with each other – that is, to share half or more of their Twitter followers – as a function of their party affiliation, controlling for the unequal size of each party as well as for several endogenous effects (namely: reciprocity, transitivity and dyadic dependence, with the weight parameters α of the terms controlling for the latter two set at 1 after testing all configurations between 0 and 1.5 at intervals of 0.1).[4]

The key result of this table is the "Same party" term, which estimates partisan homophily net of all aforementioned effects. The positive and statistically significant coefficient for partisan homophily, which can be read in similar fashion to a log-odds coefficient in a more traditional logistic regression, confirms our initial observation: in our data, sharing a large proportion of Twitter followers is approximately $\exp 1.38 \sim 4$ times more frequent among politicians who are affiliated with the same party.

We produce this result only to suggest that the kind of data exposed in this paper are also fit for analysis under different (but not unrelated) methods than those we rely on in the paper. The potential of network models seems particularly promising in that respect, although a full estimation strategy would require a more comprehensive model than the very preliminary one offered above. Furthermore, fitting a latent space model (Hoff, Raftery and Handcock, 2002) to the network under study would also require a highly efficient infrastructure to handle the high computational costs of the procedure, an option that was not available to us at the time of writing.

[4] See Hunter and Hunter and Handcock (2006) and Hunter (2007) for details on setting up these terms, which are alternatives to the combinations of k-stars and triangles to control for similar effects.

	ERGM
Edges	$-2.48\ (0.20)^{***}$
Main effect: DVD	$0.46\ (0.21)^{*}$
Main effect: DVG	$0.14\ (0.18)$
Main effect: EELV	$0.33\ (0.13)^{**}$
Main effect: FDG	$0.06\ (0.23)$
Main effect: FN	$-0.07\ (0.12)$
Main effect: MODEM	$0.10\ (0.16)$
Main effect: PRG	$0.08\ (0.13)$
Main effect: PS	$-0.06\ (0.11)$
Main effect: UDI	$0.18\ (0.10)$
Main effect: UMP	$-0.08\ (0.11)$
Same party	$1.38\ (0.10)^{***}$
Mutuality	$-3.62\ (12.90)$
GWESP	$1.09\ (0.07)^{***}$
GWDSP	$-0.50\ (0.02)^{***}$
GWD (in)	$1.60\ (0.36)^{***}$
GWD (out)	$9.68\ (2.02)^{***}$
AIC	147735.98
BIC	147936.03
Log Likelihood	-73850.99

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Exponential random graph model of the shared followers network. Alpha and decay parameters set at 1 for the geometrically weighted terms.

C Estimation of the Spatial Following Model

The core estimation strategy of the model, which is driven by Bayesian principles, works in two stages. In the first stage, the model parameters that relate to politicians are estimated through a No-U-Turn sampler, using a sufficiently large subsample of their followers and starting values that reflect the expected direction of the ideological scale.[5] In the second stage, the model parameters of all followers are estimated through a random-walk Metropolis-Hastings algorithm (Barberá, 2015, p. 80).

For the first stage of the model, we chose to estimate ideal points only for politicians who featured a clear party affiliation as well as at least one ongoing mandate, and who sent at least one tweet in the last six months. These criteria left us with 721 politicians for which we estimated ideal points based on 44,661 of their followers. In the second stage, we reverted to the initial sample of 1,008 politicians and estimated the ideal points of the 84,279 followers described in Section 2.2 of the paper.

Like Barbera (2014); Barberá (2015), we ran the first stage of the model in Stan (Stan Development Team, 2015), and the second stage in R (R Core Team, 2015). In order to assess model fit, we ran the same kind of tests as shown in (Barberá, 2015, Appendix D.3), which showed acceptable levels of convergence in the Markov chains. Heidelberger diagnostics, for instance, indicated that the distribution of the chains was non-stationary for only two politicians in our sample.

A few examples of convergence in the Markov chains are shown in Figure 2 for a few politicians and for a random user.

[5] In our case as well as in Barberá's model, these values reflect the left-right divide of political conflict. We therefore set all left-wing politicians to start at -1 and all right-wing politicians to start at $+1$ (see also Jackman, 2001).

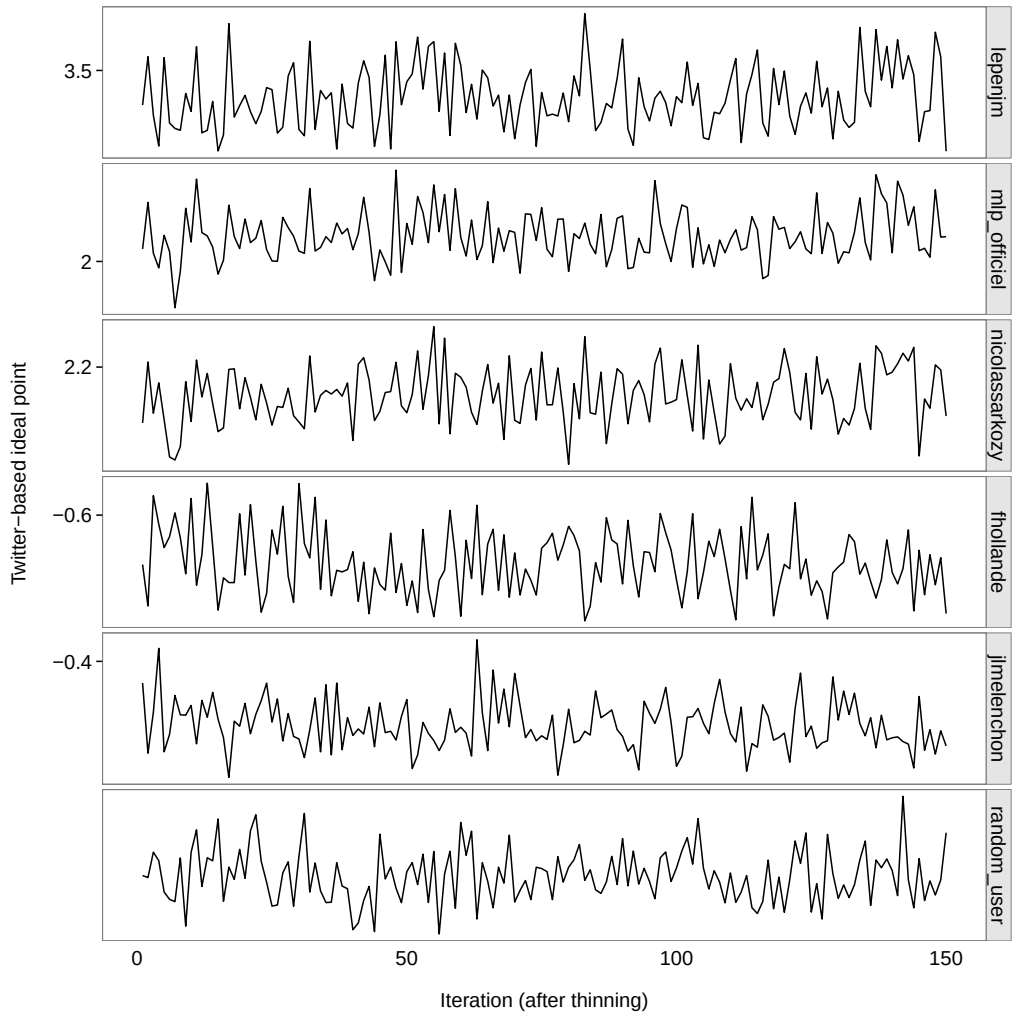

Figure 2: Traceplots of Markov chains for a selection of politicians, plus one random user.

References

Barbera, Pablo. 2014. "Replication data for: Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data.".

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political Analysis* 23(1):76–91.

Campbell, Mike and Boris New. 2000. "Fichier Prénoms 1.00.". Available at http://www. lexique.org/public/prenoms.php.

Cranmer, Skyler J and Bruce A Desmarais. 2011. "Inferential network analysis with exponential random graph models." *Political Analysis* 19(1):66–86.

Hoff, Peter D, Adrian E Raftery and Mark S Handcock. 2002. "Latent space approaches to social network analysis." *Journal of the american Statistical association* 97(460):1090–1098.

Hunter, David R. 2007. "Curved exponential family models for social networks." *Social networks* 29(2):216–230.

Hunter, David R and Mark S Handcock. 2006. "Inference in curved exponential family models for networks." *Journal of Computational and Graphical Statistics* 15(3).

Hunter, David R, Mark S Handcock, Carter T Butts, Steven M Goodreau and Martina Morris. 2008. "ergm: A package to fit, simulate and diagnose exponential-family models for networks." *Journal of Statistical Software* 24(3):nihpa54860.

IGN. 2015. "GEOFLA®.". Available at http://professionnels.ign.fr/geofla.

Insee. 2012. "Populations légales 2011 des départements et des collectivités d'outre-mer.". Available at http://www.insee.fr/fr/ppp/bases-de-donnees/recensement/populations-legales/france-departements.asp?annee=2011.

Irdes. 2014. “Population par tranche d’âge et sexe (estimations localisées de population).”. Available at https://www.data.gouv.fr/fr/datasets/population-par-tranche-d-age-et-sexe-estimations-localisees-de-population/.

Jackman, Simon. 2001. “Multidimensional analysis of roll call data via Bayesian simulation: identification, estimation, inference, and model checking.” *Political Analysis* 9(3):227–241.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *http://www.R-project.org/*

Snijders, Tom AB. 2011. “Statistical models for social networks.” *Annual Review of Sociology* 37:131–153.

Stan Development Team. 2015. “Stan Modeling Language User’s Guide and Reference Manual.”. Version 2.6.0. Available at http://mc-stan.org/.